INTERNATIONAL Journal OF DATA SCIENCE AND MACHINE LEARNING (ISSN: 2692-5141)
VOL.06 ISSUE01 (2026)

Research Article

Deep Learning Based Knowledge Assessment Systems in Education

[1]**Iskandarova Ziyoda Abdumajidovna**

[1]Senior Lecturer, Jizakh Polytechnic Institute, Uzbekistan

[2]**Iskandarova Marjona Shuxrat qizi**

[2]Bachelor's Student at Tashkent State University of Economics, Uzbekistan

Received: 10 March 2026
Revised: 22 March 2026
Accepted: 04 April 2026
Published: 25 April 2026

Doi: 10.55640/ijdsml-06-01-03
Page No: 169-173

Abstract

This study explores deep learning models for automated student knowledge assessment. Using data from 1,480 STEM students, ANN, CNN, LSTM, and a hybrid CNN-LSTM were evaluated. The hybrid model achieved highest accuracy (94.7%), outperforming baselines. Results highlight the effectiveness of combining temporal and static features for adaptive learning systems and early intervention,

Keywords:

Deep learning; knowledge assessment; student performance prediction; LSTM; convolutional neural networks; educational data mining; learning analytics; adaptive learning systems.

1. INTRODUCTION

The ability to accurately assess student knowledge and identify learning gaps is a critical challenge in modern education. Assessment serves not only as a measure of achievement but also as a diagnostic tool for guiding instruction and supporting individualized learning. However, traditional methods such as exams and graded assignments remain largely retrospective, inflexible, and subject to bias, limiting their effectiveness in capturing the dynamic nature of learning.

With the widespread adoption of Learning Management Systems (LMS), large volumes of fine-grained student interaction data have become available, including engagement logs, activity patterns, and assessment records[1]. Learning analytics research shows that such data can effectively predict academic performance, enabling the application of machine learning and deep learning methods in educational assessment.

Deep learning models offer significant advantages due to their ability to automatically learn complex patterns from raw data. In particular, LSTM networks capture temporal dependencies in sequential learning behaviors, while CNN models effectively extract structured features. Despite this potential, the use of deep learning for direct knowledge assessment remains limited and insufficiently explored [5].

This study addresses these gaps by conducting a comparative analysis of ANN, CNN, and LSTM models, and proposing a hybrid CNN-LSTM architecture that integrates static and temporal features. The results demonstrate that the hybrid model achieves superior performance, highlighting its effectiveness for automated knowledge assessment and its applicability in adaptive learning and early-warning systems.

2. METHODS

2.1. Dataset Description and Collection

The dataset was assembled from the institutional LMS (Moodle v3.11) and the student information system of a national

university over two consecutive academic semesters (September 2022 – June 2023). Inclusion criteria required active enrollment in at least one STEM course and a minimum of four weeks of recorded LMS activity, yielding a cohort of 1,480 undergraduate students across four faculties: Computer Science (n=412), Engineering (n=387), Mathematics (n=41), and Natural Sciences (n=340). The binary target variable - knowledge state - was operationalized as course completion status: students who achieved a final course grade of 60% or above were labeled as demonstrating sufficient knowledge mastery (positive class, n=774; 52.3%), while those below this threshold were classified as knowledge-deficient (negative class, n=706; 47.7%). This class distribution was considered acceptably balanced for binary classification without requiring synthetic oversampling.

Thirty features were extracted across four categories: academic performance metrics (8 features), LMS engagement indicators (11 features), behavioral sequence data (7 features), and demographic covariates (4 features). Table 1 provides a structured overview of feature categories and representative examples.

Table 1. Dataset Feature Summary

Feature Category	Feature Count	Data Type	Examples
Academic Performance	8	Continuous	GPA, quiz scores, midterm grades
LMS Engagement	11	Integer/Float	Login freq., video watch time, forum posts
Behavioral Sequence	7	Sequential	Clickstream logs, submission timestamps
Demographic	4	Categorical	Age, enrollment type, prior GPA
Total Features: 30	Total Students: 1,480	Class balance: 52.3% Pass / 47.7% Fail	

All data collection procedures were approved by the university Institutional Review Board (IRB Protocol No. 2022-EdAI-047). Student identifiers were irreversibly anonymized prior to analysis, and participation in the research was not a condition of enrollment or grading [2].

2.2. Data Preprocessing

Raw data underwent multi-stage preprocessing before model training. Missing values (3.2%) in LMS features were imputed using median values, while outliers (3.2%) were detected via IQR and clipped to the 1st–99th percentile range. All continuous features were normalized using Min-Max scaling, and categorical variables were one-hot encoded, increasing feature dimensions from 30 to 38.

For sequential models, behavioral data were structured as time-series with 16 weekly intervals, forming tensors of shape (1480 × 16 × 7). Padding and masking were applied for incomplete sequences. ANN and CNN models used static feature representations.

Four deep learning models were implemented using TensorFlow/Keras. The ANN consisted of fully connected layers with dropout regularization. The CNN applied 1D convolution to extract feature interactions. The LSTM model captured temporal dependencies in behavioral data. The proposed Hybrid CNN-LSTM combined CNN-based static feature extraction with LSTM-based temporal modeling, integrating both representations for improved classification performance.

Table 2. Model Hyperparameter Configurations

Hyperparameter	ANN	CNN	LSTM
Layers (hidden/conv)	4 dense	2 conv + 2 dense	2 LSTM + 2 dense
Units/Filters	256, 128, 64, 32	64, 128 filters	128, 64 units
Activation	ReLU / Sigmoid	ReLU / Sigmoid	tanh / Sigmoid
Optimizer	Adam (lr=0.001)	Adam (lr=0.0005)	Adam (lr=0.001)
Dropout Rate	0.30	0.25	0.35
Batch Size	64	32	32
Epochs	100 (ES*)	80 (ES*)	120 (ES*)

* ES = Early Stopping with patience = 10, monitored on validation loss. Weights restored to best epoch.

2.4. Training Protocol and Validation Strategy

All models were trained using stratified 5-fold cross-validation, ensuring that class proportions were preserved in each fold. The dataset was split at the patient level to prevent any student's data from appearing in both training and validation sets within a fold. Binary cross-entropy was used as the loss function for all models. The Adam optimizer was selected across all architectures given its adaptive learning rate properties and superior convergence behavior on sparse educational data (Kingma & Ba, 2015). Early stopping with a patience of 10 epochs was applied to prevent overfitting, with model weights restored to the epoch achieving minimum validation loss. Class weights inversely proportional to class frequency were applied during training to further mitigate the residual class imbalance. Performance metrics were computed on held-out test folds and averaged across all five splits; reported values represent mean ± standard deviation across folds. Two classical baselines - Support Vector Machine (SVM, RBF kernel, C=10) and Random Forest (500 estimators, max depth=none) - were also trained on identical feature sets for comparison purposes [3].

3. RESULTS

3.1. Comparative Model Performance

Table 3 presents the classification performance of all models on the held-out test partitions across five cross-validation folds. All deep learning architectures substantially outperformed classical machine learning baselines. Among the baselines, Random Forest (82.7%) outperformed SVM (79.4%) by 3.3 percentage points, consistent with prior literature on tabular educational data.

Table 3. Classification Performance Across All Models (Mean over 5-Fold CV)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Baseline SVM	79.4	78.1	77.6	77.8
Baseline RF	82.7	81.9	81.3	81.6
ANN	86.3	85.7	84.9	85.3
CNN	88.9	88.2	87.5	87.8
LSTM	93.1	92.6	91.8	92.2
Hybrid CNN-LSTM	94.7	94.1	93.5	93.8

Bold=best performance. Highlighted rows indicate deep learning models. Standard deviations < 1.2% across all folds for all metrics.

Among deep learning architectures, the ANN achieved 86.3% accuracy, representing a 3.6-point improvement over the best classical baseline. The CNN improved upon the ANN by 2.6 points (88.9%), demonstrating the value of convolution-based feature interaction extraction even for tabular data. The standalone LSTM achieved 93.1% accuracy, a substantial gain of 4.2 points over CNN, underscoring the importance of temporal behavioral dynamics. The proposed Hybrid CNN-LSTM model achieved the highest performance across all metrics: accuracy 94.7%, precision 94.1%, recalls 93.5% and F1-score 93.8%. The performance advantage of the hybrid model over standalone LSTM was statistically significant (paired t-test, $t = 3.41$, $p = 0.027$ at the 95% confidence level, comparing per-fold F1 scores).

3.2. Training Dynamics and Convergence

Training loss curves demonstrated that all models converged within the specified epoch budgets under early stopping constraints. The Hybrid CNN-LSTM model reached its minimum validation loss at a mean epoch of 74.4 (SD=6.1) across folds, compared to 68.2 for standalone LSTM and 52.8 for ANN, reflecting the greater representational capacity of the hybrid architecture. Validation accuracy trajectories showed no evidence of overfitting in any model after early stopping was applied: the gap between training accuracy and validation accuracy at convergence was below 2.1% for all models. The ANN exhibited the highest training-validation gap (2.0%), consistent with its susceptibility to memorizing static feature patterns; dropout regularization was sufficient to constrain this tendency without degrading generalization.

3.3. Feature Importance Analysis

A gradient-based saliency analysis was conducted on the trained Hybrid CNN-LSTM model to estimate the relative contribution of each feature category to classification decisions. Behavioral sequence features collectively accounted for the highest mean gradient magnitude (normalized importance: 0.38), followed by LMS engagement indicators (0.29), academic performance metrics (0.21), and demographic features (0.12). Within the behavioral sequence

category, assignment submission regularity (importance: 0.14) and weekly active session count (0.12) were the most discriminative individual features. This finding is consequential for deployment: it suggests that students who maintain consistent submission patterns and regular platform engagement are substantially more likely to achieve knowledge mastery, independent of raw academic performance scores.

4. DISCUSSION

4.1. Interpretation of Model Results

The superior performance of the Hybrid CNN-LSTM architecture corroborates the central hypothesis of this study: that student knowledge states are more accurately represented by the conjunction of static feature patterns and temporal behavioral dynamics than by either information stream independently. The CNN component effectively extracts interaction effects among contemporaneous features - for instance, the non-linear relationship between LMS login frequency and quiz performance that is not linearly separable - while the LSTM component captures trajectory-level signals such as an accelerating decline in engagement across the semester, which is a known precursor to course failure (Arnold & Pistilli, 2012). The concatenative fusion of these two representational streams enables the model to leverage both local feature correlations and global temporal patterns simultaneously.

The prominence of behavioral sequence features in the saliency analysis is consistent with the engagement-based theories of student retention advanced by Kuh et al. (2007), who identified behavioral engagement - characterized by sustained, effortful participation in academic activities - as the strongest predictor of persistence and achievement. The finding that assignment submission regularity outweighs raw score performance as a predictive feature is particularly noteworthy: it suggests that process-level indicators of self-regulation are more diagnostic than outcome-level metrics in early assessment contexts, where final grades are not yet available.

4.2. Comparison with Prior Literature

Table 4 contextualizes the present findings within the existing deep learning-based educational assessment literature. The Hybrid CNN-LSTM model achieves the highest reported accuracy among comparable studies, exceeding the CNN-based system of Huang et al. (2020) by 7.3 points and the LSTM-based approach of Yin et al. (2021) by 3.7 points, both of which used smaller datasets. The improvement over Yin et al. is particularly noteworthy because their dataset (n=1,050) and feature set were broadly comparable to ours, suggesting that the architectural innovation of the hybrid model - rather than dataset characteristics - accounts for the performance gain.

Table 4. Comparison with Related Studies

Study	Method	Dataset Size	Best Acc. (%)	Model Type
Romero & Ventura (2010)	Decision Tree	480	75.2	Traditional ML
Kotsiantis et al. (2004)	Naive Bayes	365	78.6	Traditional ML
Huang et al. (2020)	CNN	900	87.4	Deep Learning
Yin et al. (2021)	LSTM	1,050	91.0	Deep Learning
Present Study	Hybrid CNN-LSTM	1,480	94.7	Deep Learning

Classical machine learning studies by Romero and Ventura (2010) and Kotsiantis et al. (2004) establish the historical baseline, demonstrating that even early data-driven approaches to educational prediction achieved meaningful accuracy on substantially smaller datasets [7],[8]. The 14.7-point improvement of the Hybrid CNN-LSTM over the best classical baseline in the present study reflects the compound advantages of deeper representation learning, larger training corpora, and modern regularization techniques - advantages that were not available to earlier researchers.

4.3. Strengths and Limitations

The principal strengths of this study are its dataset scale (n = 1,480), multi-faculty heterogeneity, rigorous cross-validation design, statistical significance testing, and the systematic ablation logic embedded in the comparative model analysis. The inclusion of both classical and deep learning baselines permits an informed quantification of the marginal value of architectural complexity.

Several limitations warrant acknowledgment. First, the dataset derives from a single institution, which constrains the external validity of the findings; replication across institutions with different LMS platforms, pedagogical cultures, and student demographics is necessary before broad generalization can be claimed. Second, the binary operationalization

of knowledge master - based on a 60% threshold - simplifies a fundamentally continuous and multi-dimensional construct; future work should explore ordinal or multi-class formulations that preserve finer-grained distinctions. Third, the black-box nature of deep learning models limits interpretability for practitioners; the saliency analysis presented here offers partial mitigation, but full transparency would require integration of Explainable AI (XAI) techniques such as SHAP or LIME. Fourth, the Hawthorne effect - behavioral modification induced by awareness of being monitored - cannot be fully excluded and may have inflated engagement metrics in the dataset.

4.4. Implications for Educational Practice

The findings demonstrate strong practical relevance. An automated assessment system with 94.7% accuracy can be integrated into LMS platforms to provide real-time alerts for low-performing students, enabling early interventions such as tutoring or adjusted assessments as early as week four. Feature analysis indicates that consistent engagement and regular task completion are key factors in knowledge mastery. Additionally, the modular Hybrid CNN-LSTM architecture supports federated learning, allowing collaborative model training across institutions without sharing sensitive data, thereby addressing privacy concerns in educational AI deployment.

5. CONCLUSION

This study developed a Hybrid CNN-LSTM model for automated student knowledge assessment, achieving 94.7% accuracy and outperforming existing approaches. Results show that temporal behavioral data significantly improves predictive performance, especially when combined with static features in dual-stream architectures.

The proposed system enables real-time, scalable assessment within LMS platforms, supporting early detection of at-risk students and personalized interventions. Future work should address model interpretability and validate performance across diverse educational settings.

REFERENCES

1. Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12), pp. 267–270. ACM. https://doi.org/10.1145/2330601.2330666
2. Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. Assessment & Evaluation in Higher Education, 31(4), 399–413. https://doi.org/10.1080/02602930600679050
3. Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. Perspectives on Psychological Science, 7(5), 464–481. https://doi.org/10.1177/1745691612454304
4. Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. Review of Educational Research, 58(1), 47–77. https://doi.org/10.3102/00346543058001047
5. Huang, S., Liu, Y., & Zhao, R. (2020). Student performance prediction using convolutional neural networks on LMS data. IEEE Access, 8, 112470–112481. https://doi.org/10.1109/ACCESS.2020.2996034
6. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). arXiv:1412.6980